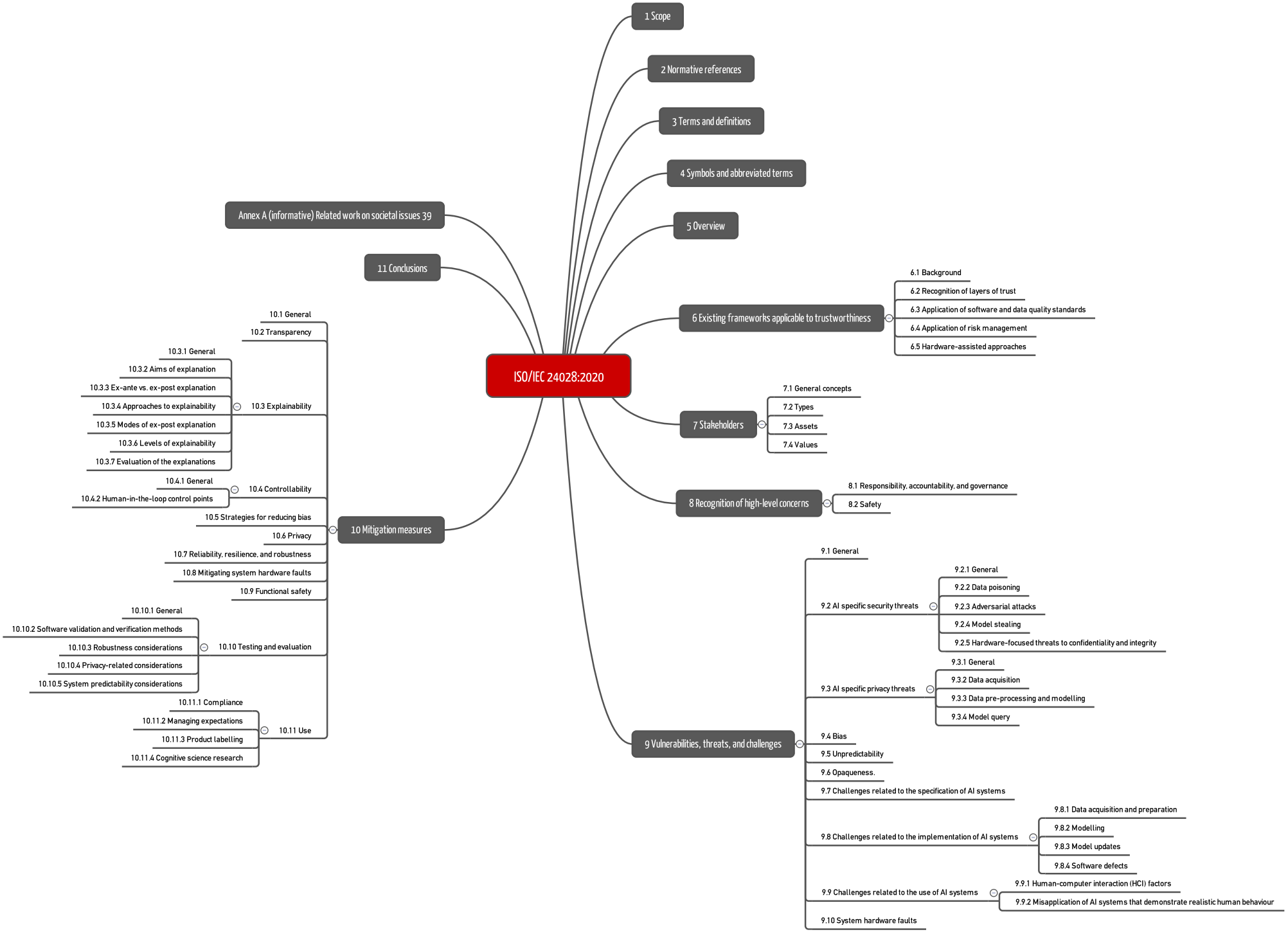
ISO/IEC 24028:2020
1 Scope
2 Normative references
3 Terms and definitions
4 Symbols and abbreviated terms
5 Overview
6 Existing frameworks applicable to trustworthiness
6.1 Background
6.2 Recognition of layers of trust
6.3 Application of software and data quality standards
6.4 Application of risk management
6.5 Hardware-assisted approaches
7 Stakeholders
7.1 General concepts
7.2 Types
7.3 Assets
7.4 Values
8 Recognition of high-level concerns
8.1 Responsibility, accountability, and governance
8.2 Safety
9 Vulnerabilities, threats, and challenges
9.1 General
9.2 AI specific security threats
9.2.1 General
9.2.2 Data poisoning
9.2.3 Adversarial attacks
9.2.4 Model stealing
9.2.5 Hardware-focused threats to confidentiality and integrity
9.3 AI specific privacy threats
9.3.1 General
9.3.2 Data acquisition
9.3.3 Data pre-processing and modelling
9.3.4 Model query
9.4 Bias
9.5 Unpredictability
9.6 Opaqueness.
9.7 Challenges related to the specification of AI systems
9.8 Challenges related to the implementation of AI systems
9.8.1 Data acquisition and preparation
9.8.2 Modelling
9.8.3 Model updates
9.8.4 Software defects
9.9 Challenges related to the use of AI systems
9.9.1 Human-computer interaction (HCI) factors
9.9.2 Misapplication of AI systems that demonstrate realistic human behaviour
9.10 System hardware faults
10 Mitigation measures
10.1 General
10.2 Transparency
10.3 Explainability
10.3.1 General
10.3.2 Aims of explanation
10.3.3 Ex-ante vs. ex-post explanation
10.3.4 Approaches to explainability
10.3.5 Modes of ex-post explanation
10.3.6 Levels of explainability
10.3.7 Evaluation of the explanations
10.4 Controllability
10.4.1 General
10.4.2 Human-in-the-loop control points
10.5 Strategies for reducing bias
10.6 Privacy
10.7 Reliability, resilience, and robustness
10.8 Mitigating system hardware faults
10.9 Functional safety
10.10 Testing and evaluation
10.10.1 General
10.10.2 Software validation and verification methods
10.10.3 Robustness considerations
10.10.4 Privacy-related considerations
10.10.5 System predictability considerations
10.11 Use
10.11.1 Compliance
10.11.2 Managing expectations
10.11.3 Product labelling
10.11.4 Cognitive science research
11 Conclusions
Annex A (informative) Related work on societal issues 39